Bayesian Hierarchical MPT Models

Theory

Daniel W. Heck

STATISTICAL MODELING in PSYCHOLOGY

FREIBURG HEIDELBERG LANDAU MANNHEIM TÜBINGEN

2018-09-11

Bayesian Hierarchical MPT Models

1. MPT models & heterogeneity
2. Hierarchical MPT models
3. Bayesian estimation with MCMC sampling
4. Advantages of MCMC
5. Application: Linking personality to MPT models

MPT models & heterogeneity

Standard MPT models assume that ...

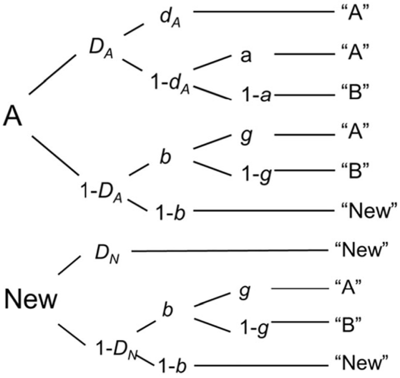
- ... people behave identically
- ... items are similarly difficult
- ... trials are independent (no order effects)
- Technical assumption
    - Fixed-effects model: Observations are "independent and identically" (i.i.d.) distributed
    - The likelihood of all observations $i = 1, \ldots, n$ is the product of the likelihood of a single observation x_i

$$p(x_1, \ldots, x_n \mid \theta) = \prod_{i=1}^{n} p(x_i \mid \theta)$$

What about real data?

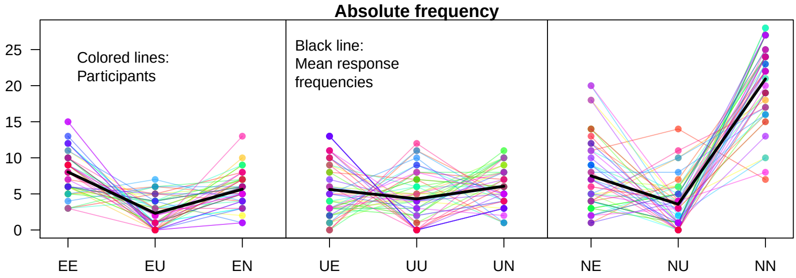
Source-Monitoring

1. Study phase: List of words from Source A and B.
2. Test phase: Is the presented item from Source A/B/New?



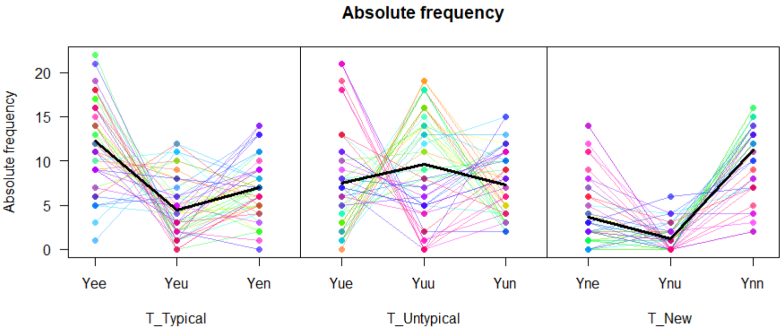
Source-monitoring task

- Distribution of individual response frequencies
- Example: Experiment on schema activation (Arnold et al., 2013)



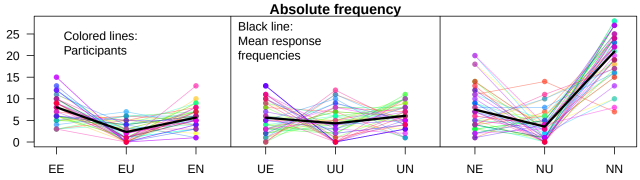
People Behave Differently

Data from a different experiment (Bayen, 2011)



- Substantial variance in the choice patterns of participants
  - Differences in memory? Response bias?
- If we fit a standard MPT model to the aggregated data, these differences are ignored (treated as random, unsystematic noise)

People Behave Differently



Heterogeneity of participants

- Response frequencies are often aggregated across subjects
  - Dependent variable: Summed individual frequencies
- However, responses are likely not i.i.d.
  - Assumption can be tested statistically (Smith & Batchelder, 2008)
- Heterogeneity may result in biased statistical inference
  - Biased point estimates if parameter are correlated
  - Over-/underestimation of confidence intervals
  - Inflated model-fit statistics

How to Handle Heterogeneity?

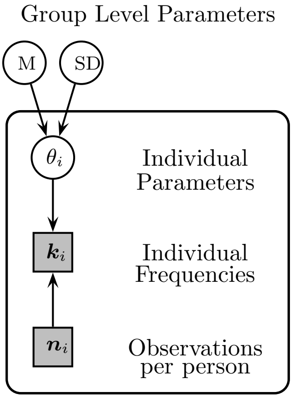
1. **Complete pooling**: Analysis of aggregated frequencies
   - Ignores differences between persons
   - High power, but possibly biased statistical inference
2. **No pooling**: A separate MPT model per person
   - Low power, parameter estimates will have a large variance
   - Often, not enough data per participant
   - Problem: How to aggregate results across models?
3. **Partial pooling**: Hierarchical model
   - Account for differences AND similarities between persons jointly
   - Provides correct statistical inferences
   - Higher efficiency than separate analysis

Note: This classification is very general and not limited to MPT models.

Hierarchical MPT models

Bayesian hierarchical MPT
(Klauer, 2010; Smith & Batchelder, 2010)

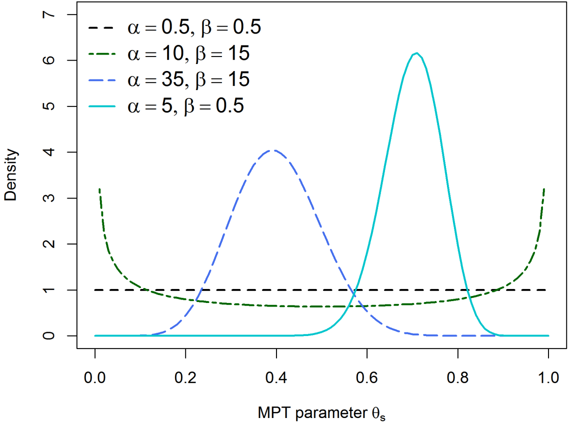
- Explicit model for participant heterogeneity
- Assumption: MPT structure holds for each person, but with different parameters!
- One parameter vector $\theta_i = (D_i, d_i, g_i, \ldots)$ per person
- On the group level, the θ_i have a specific distribution
  1. Beta-MPT: Beta distribution
  2. Latent-trait MPT: multivariate normal distribution for the probit-transformed parameters



Beta-MPT

Beta distribution

- Ideally suited to model the distribution of an MPT parameter:
  - Allows values between 0 and 1
  - Two shape parameters: α and β
- On the group level, the mean for the MPT parameter equals: $\alpha/(\alpha+\beta)$



Beta-MPT

Beta-MPT (Smith & Batchelder, 2010)

Parameters:

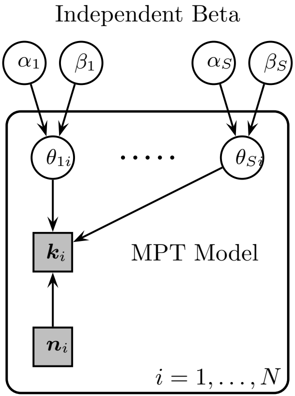
- Level-1: MPT parameters θ_{si} of person i
- Level-2: Shape parameters α_s and β_s of beta distributions

Data:

- k_i: Individual choice frequencies
- n_i: Number of responses per person

Priors:

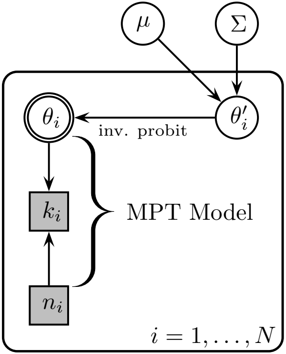
- Uniform or gamma on α_s and β_s



Latent-trait MPT (Klauer, 2010)

Parameters:

- Level-1: Person parameters are probit-transformed
  - $\theta_{si} = \Phi(\theta'_{si})$
  - Φ = cumulative density function of the standard normal
- Level-2: Probit-transformed parameters have a multivariate normal distribution
  - Mean $\boldsymbol{\mu}$ and covariance matrix $\boldsymbol{\Sigma}$ (on probit scale)



Prior distributions:

- Standard normal distributions for $\boldsymbol{\mu}$
- Scaled inverse-Wishart prior for $\boldsymbol{\Sigma}$

Transformation of MPT parameters

- We need to transform the probability parameters (d, D, ...)
- We want parameters between $(-\infty, +\infty)$ (to work with normal distributions)
- Solution: Transform parameters using the cumulative density function Φ of the standard-normal distribution (similar as in logistic regression)

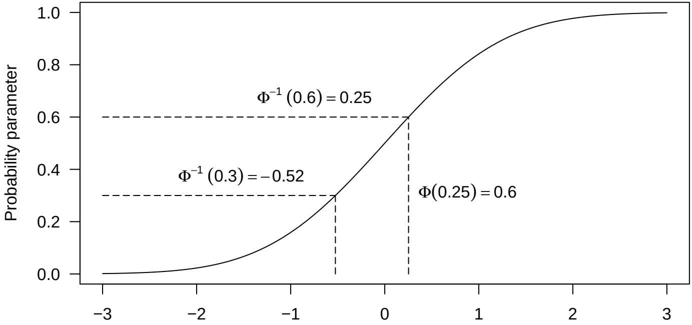
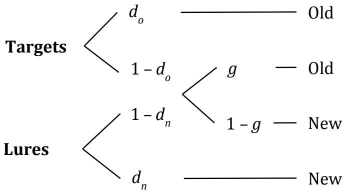


Illustration: Separate MPT Structure for each Person

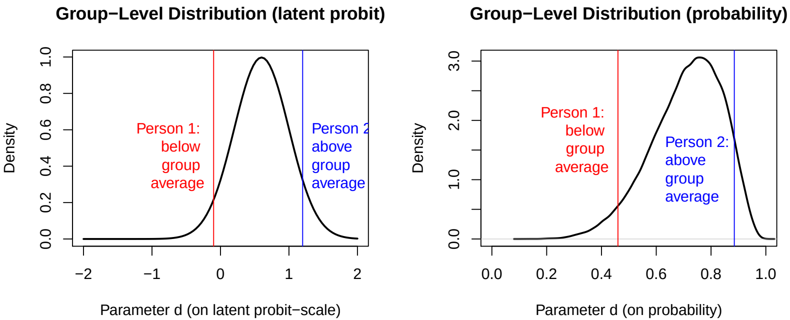
Example: 2HTM for two persons

- Probit scores for memory parameter d are: $-.10$ and 1.20
- What is the predicted probability of correct OLD responses (hits)?
- We assume symmetric and identical guessing for everybody ($g = .50$)
- **Person 1**:
  1. Transform: $d = \Phi(-.10) = .46$
  2. MPT: $P(hit) = d + (1-d)g = .46 + (1-.46).50 = .73$
- **Person 1**:
  1. Transform: $d = \Phi(1.20) = .88$
  2. MPT: $P(hit) = d + (1-d)g = .88 + (1-.88).50 = .94$



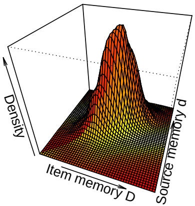
Assumption: Normal distribution of probit parameters

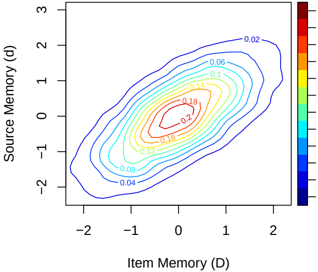
- Illustration: Normal distribution with mean $\mu_d = .80$ and standard deviation $\sigma_d = .3$
- For interpretation, it matters whether parameters are on the probit or the probability scale



Parameter correlations

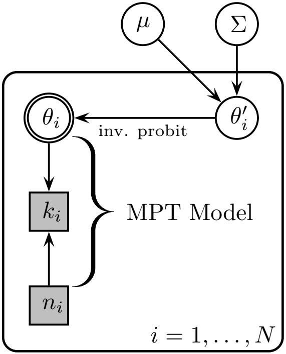
- Item and source memory might be correlated (parameters D and d)
- "The more likely I remember the item, the more likely I also remember the source."
- Solution: Assumption that the vector θ_i' with probit-transformed MPT parameters follows a *multivariate* normal distribution
- Caveat: Correlation estimates are often very unprecise and require both large number of responses and large number of participants





Core ideas of hierarchical models

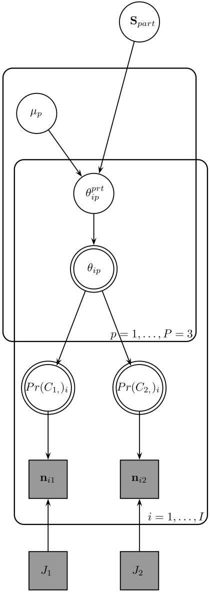
- Assume an MPT model with separate MPT parameters θ_i per person
- On the group-level, the parameters have a specific distribution
  1. Beta-MPT: Beta distribution
  2. Latent-trait MPT: multivariate normal distribution of probit-parameters with mean μ and covariance matrix Σ
  3. Other option (not discussed here): Discrete latent classes (Klauer, 2006)



Excursion: Graphical Models

Bayesian graphical models

- In publications, graphical models look more difficult
- Example: Matzke et al. (2015)
- However, most models use exactly the same ingredients



$$\mathbf{S}_{part} \sim \text{Scaled} - \text{Inverse} - \text{Wishart}(\mathbf{W}, df = P + 1, \boldsymbol{\xi}_{part})$$
$$\xi_{part_p} \sim \text{Uniform}(0, 100)$$
$$\mu_p \sim \text{Normal}(0, 1)$$

$$\boldsymbol{\theta}_i^{prt} \sim \text{Multivariate} - \text{Normal}\left((\mu_1, \ldots, \mu_P), \mathbf{S}_{part}^{-1}\right)$$
$$\theta_{ip} = \phi\left(\theta_{ip}^{prt}\right)$$

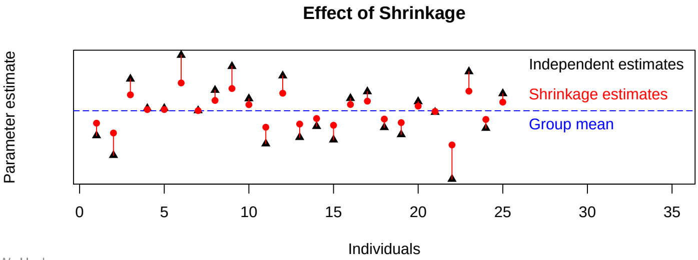
$$Pr(C_{11})_i = \theta_{i1} \times \theta_{i2}$$
$$Pr(C_{12})_i = (1 - \theta_{i1}) \times \theta_{i3}^2$$
$$Pr(C_{13})_i = (1 - \theta_{i1}) \times 2 \times \theta_{i3} \times (1 - \theta_{i3})$$
$$Pr(C_{14})_i = \theta_{i1} \times (1 - \theta_{i2}) + (1 - \theta_{i1}) \times (1 - \theta_{i3})^2$$

$$Pr(C_{21})_i = \theta_{i3}$$
$$Pr(C_{22})_i = (1 - \theta_{i3})$$

$$\mathbf{n}_{i1} \sim \text{Multinomial}\left(Pr(C_{1,})_i, J_1\right)$$
$$\mathbf{n}_{i2} \sim \text{Multinomial}\left(Pr(C_{2,})_i, J_2\right)$$

Benefits of hierarchical MPT models

- Avoid aggregation biases
- “Shrinkage” of parameter estimates
  - Parameter estimates for each person are closer together compared to fitting each person separately
  - Hence, extreme estimates are less likely
  - Overall, this ensures that parameter estimates are closer to the true values on average
- The basic idea of hierarchical models can easily applied to any other model
  1. Assume that model holds for each person
  2. Specificy group-level distribution of parameters across persons



Bayesian estimation with MCMC

Fitting Hierarchical MPT Models

Parameter estimation

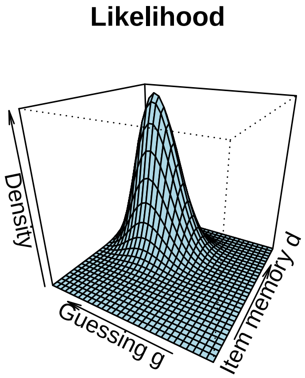
- How can we actually fit such models?
- Which are the “best” parameters given the data?
  - Standard MPT models: Maximum likelihood estimation
  - Not an option for hierarchical models (intractable likelihood function due to high-dimensional integrals)

Solution

- Hierarchical models are often fitted using Bayesian statistics

Maximum Likelihood

- Logic of parameter estimation with maximum-likelihood
  1. Define likelihood function $p(x \mid \theta)$
  2. Find parameters θ that maximize f
- Interpretation: "The estimator $\hat{\theta}$ has the highest likelihood."
- Computational solution: Algorithm searches for the "top of the mountain"

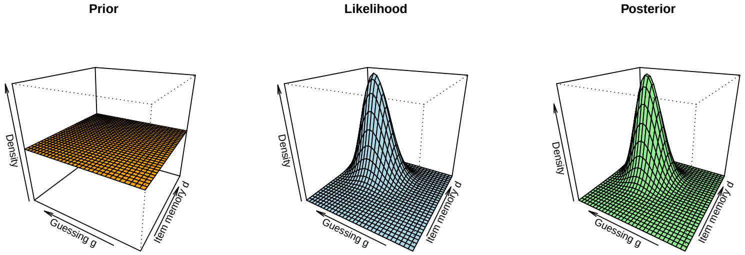


Bayesian Estimation

- Logic of Bayesian parameter estimation
  1. Define likelihood $p(x \mid \theta)$ and prior distribution $p(\theta)$
  2. Derive the posterior distribution of the parameters via Bayes' theorem:

$$p(\theta \mid x) = \frac{p(x \mid \theta) p(\theta)}{p(x)}$$

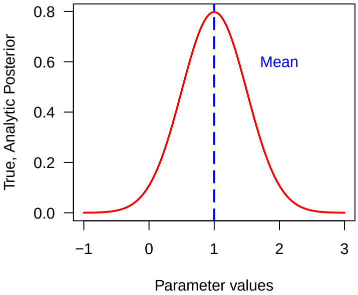
- Interpretation: "What have we learned about the parameters θ given the data x?"

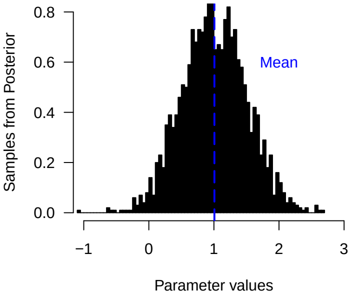


Bayesian Estimation

- Problem: We need to work with the posterior function $p(\theta \mid x)$
  - What is the mean/mode/95% credibility interval of θ?
  - Often, this is analytically not tractable
- Solution: We draw random samples from the posterior distribution
  - Logic: It is easier to draw conclusions from these random samples than deriving solutions for the analytical posterior (which is a function!)
  - Example: Computing the mean of a normal distribution requires to solve:

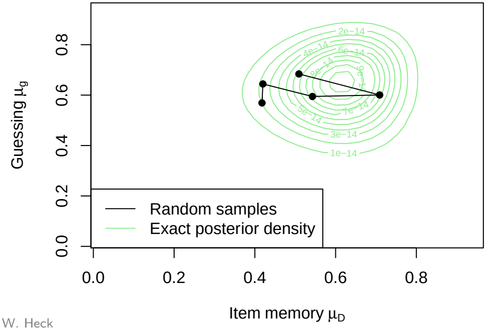
$$E[X] = \int_{-\infty}^{+\infty} x \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right) dx$$





Markov Chain Monte Carlo (MCMC) Sampling

- We draw random samples of the posterior distribution for *all* parameters (individual and group level)
- Simplified example for two parameters of the 2HTM:
  1. First sample: $(\mu_d = .6, \mu_g = .71, \sigma_d = .1, \ldots, d_1 = .8, \ldots)$
  2. Second sample: $(\mu_d = .73, \mu_g = .5, \sigma_d = .12, \ldots, d_1 = .67, \ldots)$
  3. $\ldots$
- Once we have the samples:
  - Compute the mean of the posterior samples to get parameter estimates



Markov chain Monte Carlo (MCMC)

- General method to draw posterior samples
- In a hierarchical model, there are many (!) parameters
  - Group-level means and covariances, person parameters, ...
  - Intuitively, this method moves around and searches for parameter values with high posterior density
- There are software packages that draw random samples for many models of interest
  - JAGS, WinBUGS, OpenBUGS, Stan, ...

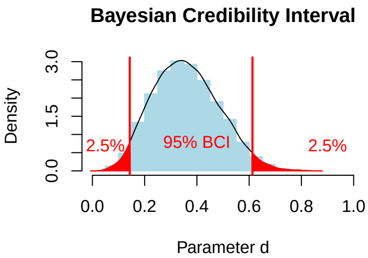
Summary of Bayesian estimation

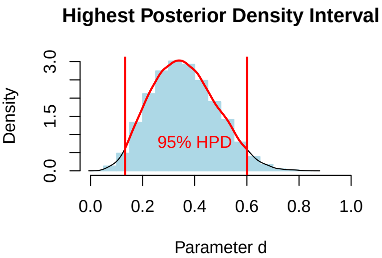
1. Develop a model (=> psychological theory, multiTree)
2. Get posterior (MCMC) samples (JAGS, TreeBUGS)
3. Summarize these samples (e.g., mean of group-level parameters μ_D, μ_g, ...)

Advantages of MCMC

Advantages of MCMC sampling

- Theoretical:
  - No asymptotic assumptions
  - Maximum likelihood: requires a sufficient number of observations
- Practical: It is easy to quantify uncertainty
  - Bayesian credibility interval (BCI): What are the 2.5%- and 97.5%-quantiles of the parameter values?
  - Highest posterior density interval (HPD or HDI): What are the 95% most plausible parameter values?
  - For probability parameters, these intervals will always be in the interval $[0, 1]$



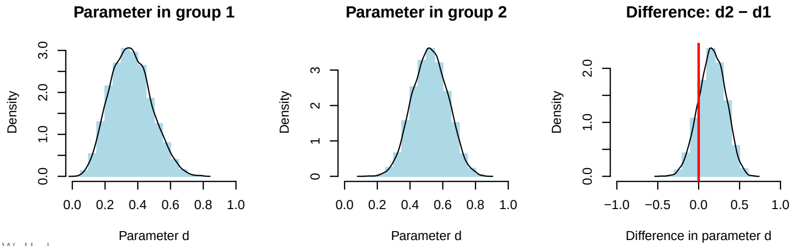


Advantages of MCMC: Transformed Parameters

- Often, we are interested in parameter/group comparisons
  - Example: Do healthy controls vs. schizophrenics differ in memory?
  - Test: Does the group-mean parameter μ_D differ?
- Based on MCMC samples, we can directly estimate functions of the parameters

MCMC estimation of transformed parameters

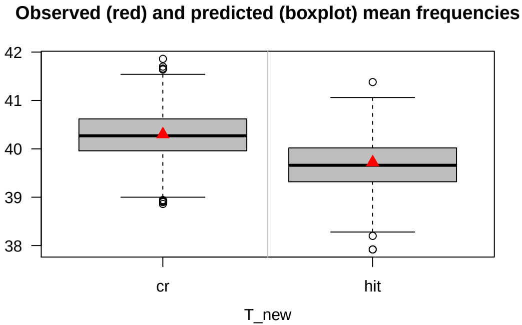
1. Draw MCMC samples
2. Compute transformed parameters for all samples
   - Example: $\delta^{(t)} = \theta_1^{(t)} - \theta_2^{(t)}$
3. Summarize the new values



Advantages of MCMC: Model Fit

Does the model fit the data?

- Graphical comparison: observed vs. predicted frequencies
  - Use posterior samples of the MPT parameters to sample new data (= posterior predictive)
  - Compare whether these predicted data (boxplot) are in line with the observations (red points)



How to quantify model fit for MPT models?

- Test statistic similar to Pearson's X^2 statistic (Klauer, 2010)
  - T1 statistic: Mean structure of frequences
  - T2 statistic: Covariance matrix of frequencies
- Posterior predictive p-value (PPP) measures model fit:
  1. Compute T1 for the observed data
  2. Compute T1 for the posterior predicted data
  3. PPP $=$ probability that T1(predicted) is larger than T1(observed)
- Ideally, PPP should be around .50

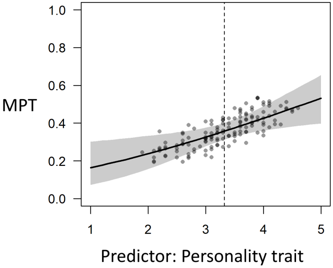
Application

Linking Personality Traits to MPT Parameters

Interindividual differences

- Personality
  as a predictor for MPT parameters
- Statistical approach in latent-trait MPT: Similar to logistic regression

$$p_i = \Phi(\mu + \boxed{\beta \cdot x_i} + \delta_i)$$



Cognitive Psychometrics

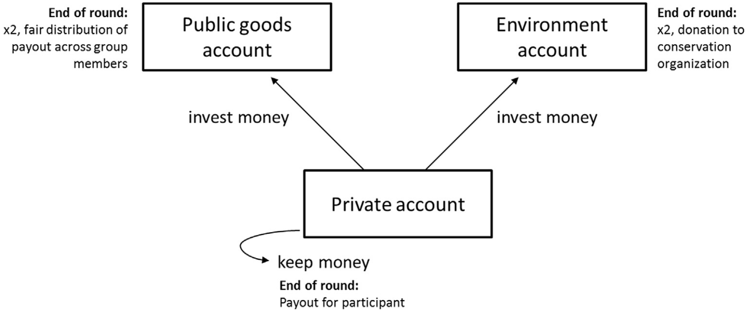
- Talk: "Bayesian Hierarchical Multinomial Processing Tree Models: A General Framework for Cognitive Psychometrics"
- Wednesday, 10:45–11:30
- Room: H34

Example: Linking personality to MPT parameters

- "Which is the greater good? A social dilemma paradigm disentangling environmentalism and cooperation"
  - Klein, Hilbig, & Heck (2017). *Journal of Environmental Psychology*)
- Research question: How can we distinguish between 3 types of behavior?
  - Pro-environmental behavior
  - Pro-social behavior
  - Selfish behavior

Application: The Greater Good Game

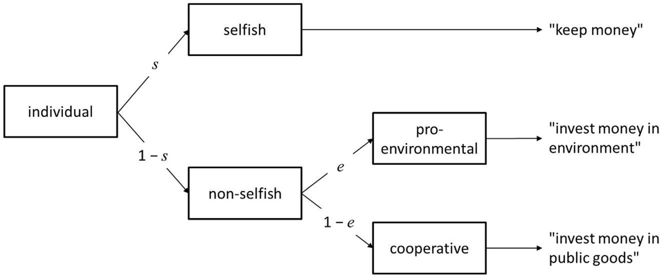
- Greater Good Game
  - Participants decide whether to keep the money for themselves or contribute it to either a public goods or an environment account.
  - Important: Participants are forced to decide between the group and the environment!
  - The game is a variant of a nested public goods game



Application: MPT Model

MPT model for the Greater Good Game

- s = probability of selfish behavior
- e = probability of pro-environmental behavior



Results

- Honesty Humility (= sincerity, fairness) is associated with less selfish behavior
- Selfish behavior decreases from 33.4% to 13.9% for participants $-1/+1$ SD on Honesty Humility

Hierarchical MPT Models

- Individual level
  - Assume a separate MPT model for each person
- Group level
  - Beta-MPT: Beta distribution of person parameters
  - Latent-trait MPT: Normal distribution of probit-transformed parameters
- Bayesian model fitting: Drawing posterior samples

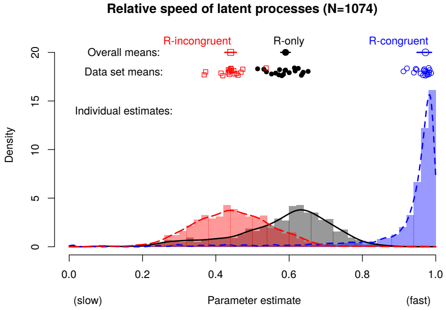
Appenidx & References

Appendix A: Meta-Analysis of Raw Data

- Linking process and measurement models of recognition-based decisions (Heck & Erdfelder, 2017, PsychReview)
- Reanalysis of about 400,000 decisions
  - 3-level hierarchical latent-trait MPT:

$$\theta_{sij} = \Phi(\mu_s + \boxed{\xi_{sj}} + \delta_{si})$$

  - Overall mean of MPT parameters (μ_s)
  - Participants nested in studies (random effect: ξ_{sj})
  - Responses nested in participants (random effect: δ_{si})



Appendix B: Large-Scale Reanalysis

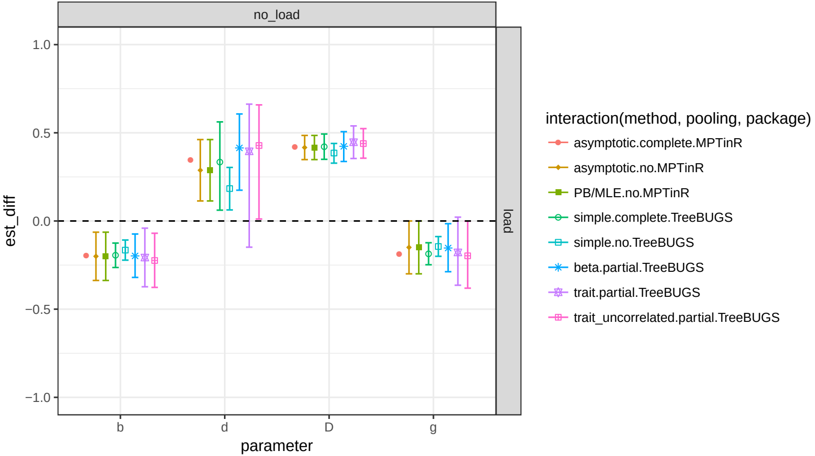
Open questions:

- How much do results actually differ between different MPT versions?
- Which MPT version should be used in practice?

Large-scale reanalysis project

- Network of MPT researchers (organized by Beatrice Kuhlmann & Julia Groß)
- Reanalysis of existing data sets to compare:
  - Fixed-effects vs. hierarchical
  - Maximum-likelihood vs. Bayes
  - Different hierarchical level-2 structures (beta, multiv. normal, independent univ. normal)
- Software: "A multiverse pipeline for MPT models"
  - Maximum likelihood: `MPTinR` (Henrik Singmann)
  - Bayes: `TreeBUGS`
  - Available at: https://github.com/mpt-network/MPTmultiverse

- Source-monitoring model (data by Bayen & Kuhlmann, 2011)
- Plot: Difference in parameters across two groups



References

- TreeBUGS and a simple introduction to hierarchical MPT models
  - Heck, D. W., Arnold, N. R., & Arnold, D. (in press). TreeBUGS: An R package for hierarchical multinomial-processing-tree modeling. Behavior Research Methods. https://doi.org/10.3758/s13428-017-0869-7
- The latent-trait model (very technical)
  - Klauer, K. C. (2010). Hierarchical multinomial processing tree models: A latent-trait approach. Psychometrika, 75, 70–98. https://doi.org/10.1007/s11336-009-9141-0
- The latent-trait model with crossed-random effects and a JAGS implementation
  - Matzke, D., Dolan, C. V., Batchelder, W. H., & Wagenmakers, E.-J. (2015). Bayesian estimation of multinomial processing tree models with heterogeneity in participants and items. Psychometrika, 80, 205–235. https://doi.org/10.1007/s11336-013-9374-9

References

- Alternative hierarchical group structure of parameters
  - Smith, J. B., & Batchelder, W. H. (2010). Beta-MPT: Multinomial processing tree models for addressing individual differences. Journal of Mathematical Psychology, 54, 167–183. https://doi.org/10.1016/j.jmp.2009.06.007
- Benefits of hierarchical cognitive models
  - Lee, M. D. (2011). How cognitive modeling can benefit from hierarchical Bayesian models. Journal of Mathematical Psychology, 55(1), 1–7. https://doi.org/10.1016/j.jmp.2010.08.013